CAN WE PREDICT F'2 BY MEANS OF A SIMILARITY MEASURE ?

Denis TUFFELLI & Haiyan YE

Laboratoire de la Communication Parlée (ICP Unité Associée au CNRS)
INPG-ENSERG 46, Avenue Félix Viallet
38031 GRENOBLE CEDEX, France

ABSTRACT

The prediction of F'2 is an important aspect for vowel perception. Several prediction models have been proposed in the recent years. In these studies, relationships with the Center of Gravity (in particular with broad band integration) are important. In this paper we propose a new approach for the prediction of F'2 by means of measures of similarity and/or dissimilarity. Several algorithms have been tried including an integration by a critical distance dynamic programming (CDP) and a critical distance transformation (CDT). The evaluation tests are carried out with two kinds of data: vowels formants frequencies and synthetic vowels. The results show that the CDT with a simple euclidean distance give good results. This transformation could retain the phonetic qualities of a sound and give us a good spectral representation for a speech recognition system.

INTRODUCTION

Previous works have underlined two interesting phenomenons: center of gravity of spectral peaks and the F'2 of vowels /1-4/.

The center of gravity (CG) found at Leningrad, is the rough estimate by listeners of two formants F1 and F2 with one formant Fv. In short, a listener hears a first sound made up by F1 and F2 then he is asked to vary the Fv frequency of a second sound in order to find the "best" Fv. If the gap between F1 and F2 is less than 3.5 Barks (so called critical distance) the listener adjusts Fv between F1 and F2 (near the center of gravity), else F1 or F2 is found as the best value for Fv.

With F'2 the principle is similar but the first sound is made up by four formants F1,F2,F3,F4 and the second sound by two formants F1,Fv. The best Fv is called F'2 (effective second formant).

We found that it is hard to simulate both these experiments with a machine "operator" instead of a listener. It is the aim of this paper to describe the machine operators we used.

With CG as a first attempt, we could try t compare two sounds by an euclidean distance D on tw spectra Rf(F1,F2) and Tf(Fv) (with pure peaks at F1 F2 and Fv). The best Fv^* frequency could be define as:

$$D(Fv^*) = \min_{Fv} D(Fv)$$

with any gap, between F1 and F2, the result is F1 o F2. So by extrapolation on formants with smal bandwiths, we consider that this result is no correct.

Different hypothesis can be made. For instanc with F'2 previous works /5/ have lead to thre hypothesis for F'2 perception:

1) after one broad band integration a featur extractor detects F'2 as a parameter for vowe identification.

2) F'2 is a by product of a classificatio process.

3) F'2 is a by product of a similarity o dissimilarity evaluation of the auditory system.

In this paper we have chosen the 3r hypothesis. We will use a distance measure betwee two sounds S1 and S2, so called afterwards D(S1,S2). The basis parameters will be two spectra Rf and T with N components on a mel scale.

COMPLEMENTARY TESTS

Of course we measured the machine operator quality by the obtained values on F'2 and CG, but w thought that it was not sufficient. We used tw others tests:

The first complementary test is the followin one. Listeners were asked to determine th boundaries between vowels pair. That is, fo instance for one vowel pair V(F1,F2,F3,F4) an V'(F'1,F'2,F'3,F'4) we generated intermediate sound Ui by formants interpolations. Therefore thes formants had the values:

$$b_iF1 + (1-b_i)F'1$$
$$b_iF2 + (1-b_i)F'2$$
$$\dots \text{ etc.}$$

Where b_i is a value between 0 and 1.

Then we determined with the listeners the i value which gives the maximal ambiguity between th two vowels V and V'. We compared the obtained value i^* with those given by the operators (for one pair the machine boundary i is roughly determined by the equality D(V,Ui)=D(Ui,V').

The second complementary test is the following checking: with two inputs spectra Rf(F1,F2) and Tf(F1,Fv), the best Fv frequency must be equal to F2 (if the amplitudes are the same). This condition seems obvious but it is not necessarily verified with dynamic programming based algorithms.

ALGORITHMS PRINCIPLES

A Critical Distance Dynamic Programming Algorithm (CDP)

The classic distance measures compare two spectra ,component by component, at the same frequency. If we draw a graph with Rf on the x axis and Tf on the y axis, in this case the followed path is the diagonal. Some people proposed that any kind of paths should be possible /6-8/. They used dynamic programming to get the best path. Each graph node (with coordinates x,y) had a weight which was computed by an elementary distance d(Rx,Ty). The obtained results were not very satisfactory. Following this idea, we propose here that an horizontal or vertical segment is the result of an "integration" (Fig.1). The maximal lenght of such a segment is 3.5 Barks, that is the maximal warping allowed.

To get the best path we try to get a maximum of an inter-spectrum correlation Cxy which is weighted by a distortion term. This term measures the distance to the diagonal. Cxy is a value tied to each point (x,y) and is defined as:

$$Cxy = Rx*Ty*(1 - ((x-y)/alpha)^2)$$

We can see that the distortion term:

$$(1-((x-y)/alpha)^2)$$

is maximum on the diagonal and becomes small when |x-y| tends towards alpha.

If we take pure peaks at frequencies F1, F2 and Fv, during a "machine experiment" of the center of gravity, the best path comes through the horizontal segment (F1,Fv) (F2,Fv) with $Fv=(F1R_{F1}+F2R_{F2})/(R_{F1}+R_{F2})$. Therefore Fv is the mathematical center of gravity. From the best path we can get Fv. Here we don't compute a real distance.

A Critical Distance Transformation (CDT)

The previous technique is an awfully time consuming one. It is the result of a particular interpretation of the human experiments from an algorithmic point of view. Another interpretation can be that the human results are the consequences of a particular preprocessing. We applied it to a spectral preprocessing we are going to describe. Then the distance to use becomes simple (for instance euclidean type on the CDT preprocessed spectra).

Starting from a spectrum Sf, we get the transformed spectrum Sf^* from the formula:

$$S^*f = \max_x \sum_{x}^{x+Cd} Sx (1 - ((f-x)/alpha)^2) \quad (1)$$

with |f-x| smaller than alpha.

The distance to use between two spectra Rf, Tf is:

$$D(R,T) = \sum_{i=1}^{N} || Rf^* -Tf^* ||$$

If we take a spectrum Sf which consists of two pure peaks at frequencies F1 and F2 with amplitudes a1 and a2, we have (providing that F1 and F2 are not too far and Fv belongs to some frequency range):

$$S^*f = a1(1-((F1-f)/alpha)^2) + a2(1-((F2-f)/alpha)^2)$$

S^*f is a parabola with a maximum at the frequency Fv which is the mathematical center of gravity:

$$Fv = (a1F1+a2F2)/(a1+a2)$$

Of course one can find always the center of gravity with more than two spectral peaks between F1 and F2. At last one can demonstrate that the resulting distance is almost linear, in some particular cases, with the gap between spectral peaks.

The maximum in the formula (1) does not seem necessary, but without it we have a too broad integration in our first experiments. Others experiments are necessary.

RESULTS

The alpha and Cd parameters, of the previous section, were tuned to get the best results. Generally the tuning is very difficult because there are sharp discontinuities when two formants are integrated or not.

Moreover we made some modifications on the previous formulas to improve the results, for instance with the CDT algorithm:

- when we worked on real signals, the input LPC spectra were too "soft" for this technique, we had to add a formants enhancement procedure.
- The parabolic terms, like $(1-((f-x)/alpha)^2)$, had to be slightly modified. We improved the continuity of the curves and we introduced a slight dissymmetry in the computation.
- We introduced also a slope term in the expression of the distance D.

The CDP algorithm was very difficult to tune. We had to introduce sizable modifications (Moreover the second complementary test is not verified).

Results with spectral peaks

We use the results of a previous experiment which has been carried out essentially with nine swedish vowels (in Hz) /2/:

	F1	F2	F3	F4	F'2(human)
u	310	730	2250	3300	730
o	400	710	2460	3150	720
O	360	1690	2200	3390	1720
a	580	940	2480	3290	960
y	255	1930	2420	3300	2010
U	280	1630	2140	3310	1730
e	375	2060	2560	3400	2370
ae	605	1550	2450	3400	1960
i	255	2065	2960	3400	3210

The estimated F'2 by CDP and CDT (with peaks as inputs) are as following (in Hz):

	F'2(CDP)	E_{abs}	F'2(CDT)	E_{abs}
u	742	0.08	740	0.06
o	725	0.04	720	-0.01
O	1830	0.4	1880	0.58
a	949	-0.07	950	-0.05
y	2084	0.24	2200	0.58
U	1774	0.16	1800	0.24
e	2216	-0.45	2340	-0.10
ae	1938	-0.07	1770	-0.66
i	3097	-0.24	2980	-0.50
E_{tr}		0.19		0.31
E_m		-0.45		-0.66

Where E_{abs} is the absolute error in Bark.
E_{tr} is the total mean absolute error in Bark.
E_m is maximal error.

We obtained also similar results with the center of gravity. We found that with our methods it is possible to get good results for F'2 or center of gravity. But it is very difficult to obtain good results for both of these experiments (F'2 and CG). More some F'2 values from previous experiments (Bladon & Carlson) seem incompatible and may be these values are also language depending. At last the employed energies are not sufficiently well defined and are difficult to reproduce. For further work it is necessary to get more accurate values from human experiments.

Results with synthetic sounds

We have tested the CDT algorithm with synthetic vowels by using an another criterion (first previous complementary criterion), because this method seems to us promising. This criterion is a correlation coefficient with respect to human phonetic judgements. The comparison procedure is described in /10/.

In Fig.2 one can find a LPC spectrum and a CDT filtered spectrum of the same signal. We can see that higher spectral components are well integrated. With input FFT spectra, the results are similar.

The correlation coefficient of CDT euclidean distance with respect to human phonetic judgement are between 0.87(test X) and 0.895(test ABX) for the 11 french vowels. This means that CDT has retained a great deal of phonetic information. By comparison the Itakura distance obtained, with this method, the values 0.88(test X), 0.91(test ABX). As an example one can find on figure 3, the distance behaviour between two vowels.

CONCLUSIONS

This study is just a try to predict some perceptual parameters (Center of Gravity and F'2) by means of a measure of similarity. These methods can give us a precise estimation of these parameters. Through this study, we can see that modelization of perceptual phenomena can be conducted by different ways.

The advantage of our methods is that a priori knowledges about formants are not necessary. So they can be applied to any spectra, even consonants. The application of these methods to speech recognition is more delicate and is to be tested. The CDP algorithm does not seem well adapted for that.

The phenomena of F'2 is very closely linked with human phonetic judgement. A preprocessing (similar to CDT) which can not only retain but also enhance F'2 parameter will be certainly a better and robust preprocessing for speech recognition.

ACKNOWLEDGEMENT

The authors wish to thank J.L. Schwartz and M. Mantakas for the fruitfull discussions about speech perception.

REFERENCES:

/1/. L.A. CHISTOVICH, "Central Auditory Processing of Peripheral Vowel Spectra" J. Acoust. Soc. Am. 77(3), 789-805, 1985

/2/. R. CARLSON, B. GRANSTROM, G. FANT, "Some Studies Concerning Perception of Isolated Vowel" Speech Transmission Lab. QPSR 2-3, 1970

/3/. A. BLADON, "Two-formants Model of Vowel Perception: Shortcoming and Enhancement" Speech Communication 2, 305-313, 1983

/4/. K.K. PALIWAL, D. LINDSAY, W.A. AINSWORTH, "A Study of Two-formant Models for Vowel Identification", Speech Commun. 2, 295-304, 1983

/57. J.L. SCHWARTZ & P. ESCUDIER, "Le Système Auditif Humain Comprend-il un Mécanisme d'Integration à Large Bande ?" 14 JEP, Aix-en-Provence 1986

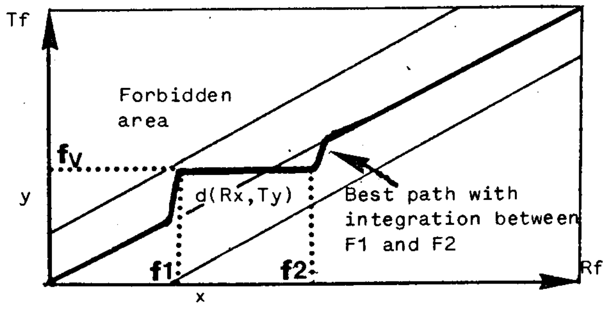
/6/. H. MATSUMOTO & H. WAKITA, "Frequency Warping for Nonuniform Talker Normalization" Int. Conf. Acous. Speech Sig. Proc., pp566-569, 1979

/7/. K.K. PALIWAL, W.A. AINSWORTH, "Dynamic Frequency Warping for Speaker Adaptation in Automatic Speech Recognition" Journals of Phonetics 13, 123-134, 1985

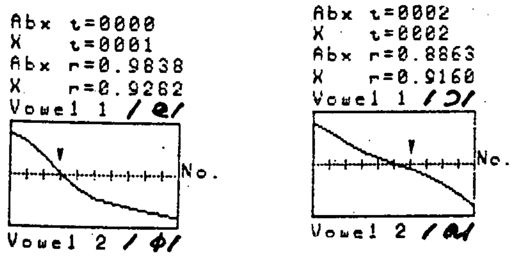
/8/. M. BLOMBERG & K. ELENIUS, "Nonlinear Frequency Warping for Speech Recognition" Int. Conf. Acous. Speech Sig. Proc., 49.2, 1985

/9/. L.A. CHISTOVICH, V.V. LUBLINSKAYA, "The Center of Gravity Effect in Vowel Spectra and Critical Distance Between the Formants: Psychoacoustical Study of the Perception of Vowel-like Stimuli" Hearing Reseach 1, pp 185-195, 1979

/10/. D. TUFFELLI & H. YE "Distortion Measures Evaluation Using Synthetic Sounds and Human's Perception" Montréal Symposium on speech recognition, (1986)



A path with "integration"
Fig.1.



Two examples of distance behaviour (with a CDT preprocessing) between a vowel pairs (V and V'). A comparison can be made with human phonetic judgements (cf /10/ for details). Here we have two vowel pairs /e/-/Φ/ and /ɔ/-/a/. t is an error number with respect to the perceptual boundary. r is a correlation coefficient between distances and perceptual data. Abx and X are two kinds of experiments. The zero crossing points are the discrimination points of the distance D. The arrows are human perception boundary. On the x-axis the numbers of the intermediate sounds Ui (from V on the left to V' on the right). On the y-axis the value D(Ui,V')-D(V,Ui).

Fig.3.

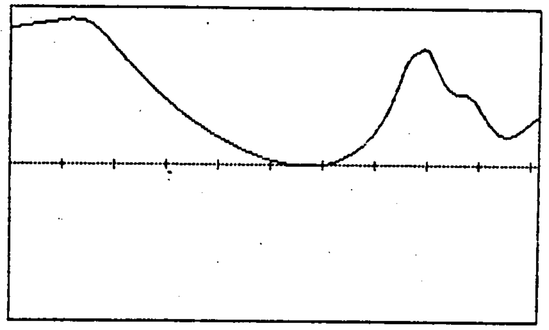


Fig.2a. LPC spectrum of a /i/
y axis: dB, x axis: Mel scale

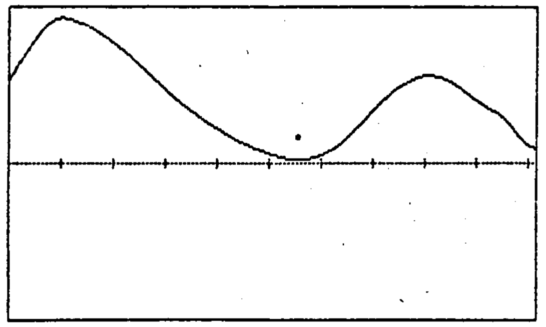


Fig.2b. CDT spectrum of the same /i/
y axis: dB, x axis: Mel scale